Título do relatório de estágio que identifique a atividade realizada

Arthur de Melo Rezende

1

Abstract. *This report aims to analyze the evolution of parallel programming techniques applied to neural networks for face recognition using CUDA, comparing the results obtained in a 2013 study with those from a modern Google Colab environment utilizing a T4 GPU. We replicated the techniques in CUDA specifically and present a performance comparison with the results from the reference paper.*

Resumo. *Este relatório tem como objetivo analisar a evolução das técnicas de programação paralela aplicadas a redes neurais para reconhecimento facial usando CUDA, comparando os resultados obtidos em um estudo de 2013 com os resultados de um ambiente moderno do Google Colab utilizando uma GPU T4. Replicamos as técnicas em CUDA especificamente e apresentamos uma comparação de desempenho com os resultados do artigo de referência.*

1. Introdução

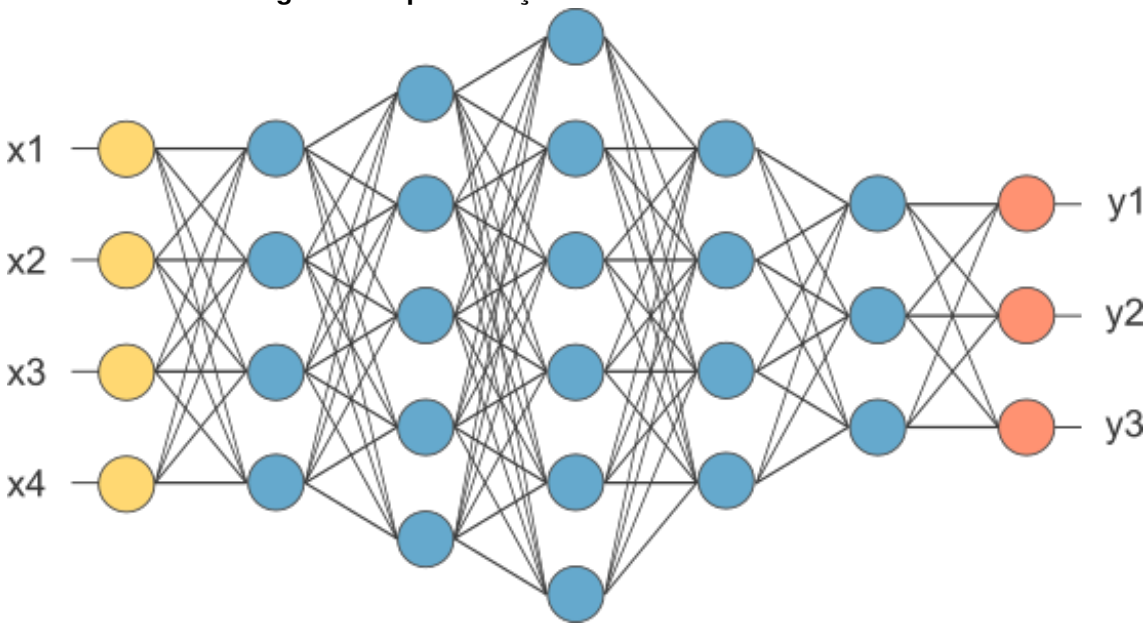
O reconhecimento facial utilizando redes neurais é utilizado em várias aplicações, entretanto esse processo conseome muito tempo, porém aevolução das técnicas de programação paralela, especialmente com o uso de GPUs, vem sendo uma ótima alternativa para a melhoria do desempenho dessas redes neurais. Este relatório visa replicar e comparar as técnicas apresentadas no artigo de 2013 "Multicore and GPU Parallelization of Neural Networks for Face Recognition" [2] com implementações modernas em um ambiente Google Colab utilizando uma GPU T4.

2. Fundamentação Teórica

Uma rede neural pode ser dito como sistemas computacionais inspirados no funcionamento do cérebro humano, capazes de aprender e generalizar padrões a partir de dados. Elas são compostas por camadas de neurônios artificiais, onde cada neurônio processa uma entrada e transmite o resultado para os neurônios da próxima camada. As redes neurais têm sido amplamente utilizadas em tarefas de reconhecimento de padrões, como reconhecimento de voz, processamento de linguagem natural e reconhecimento facial, como dito em [1]. A imagem a abaixo mostra uma representação visual de uma rede neural simples.

Umas das formas mais eficientes em melhorar uma rede neural é com a utilização de técnicas de paralelização. A paralelização é uma técnica essencial para acelerar o treinamento e a inferência de redes neurais, permitindo que múltiplas operações sejam realizadas simultaneamente. Em uma rede neural, muitas operações, como a multiplicação de matrizes e a aplicação de funções de ativação, podem ser executadas em paralelo [3], tornando a utilização de GPUs altamente eficaz.

Figure 1. Representação visual de uma rede neural



Fonte: https://www2.decom.ufop.br/imobilis/fundamentos-de-redes-neurais/

A empresa NVIDIA é uma pioneira nessa expansão na criação de novas GPUs potentes capazes de expandir o uso da paralelização nas GPUs, e isso vem sendo possível com a criação da linguagem de programação CUDA. Ela permite que os desenvolvedores utilizem a potência das GPUs para computação de propósito geral (GPGPU). CUDA facilita a criação de programas que podem executar milhares de threads em paralelo, aproveitando ao máximo a arquitetura de hardware das GPUs.

3. Ferramentas utilizadas

O algoritmo utilizado para reconhecimento facial foi implementado utilizando CUDA no ambiente Google Colab. As principais etapas incluem: Carregamento e pré-processamento dos dados de imagem; Implementação de uma rede neural para reconhecimento facial; Treinamento da rede utilizando paralelização em GPU; Avaliação do desempenho e comparação com os resultados do artigo de 2013.

Os experimentos foram realizados no Google Colab, utilizando uma GPU Tesla T4. O ambiente de teste foi configurado com CUDA 10.1 e PyCUDA para a implementação da rede neural. O código, juntamente com o dataset utilizado, pode ser encontrado no seguinte link do github: https://github.com/ArthurRLA/Atividade_Final_Computa-o_Paralela

4. Experimentos feitos

Os resultados dos experimentos mostraram uma melhoria significativa no desempenho devido à evolução das GPUs e das técnicas de paralelização. Os testes foram feitos com imagens com resolução 32x30 com 100 epochs de treinamento e 128x120 com 20 epochs de treinamento.

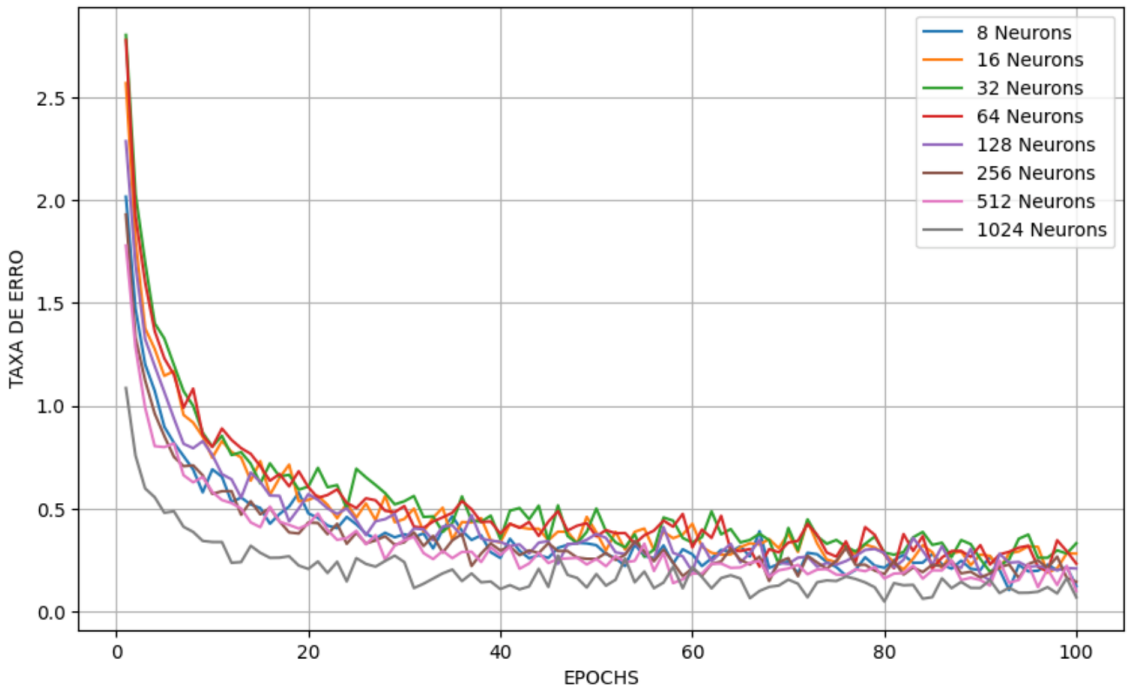


Figure 2. Taxa de erros com resolução 32x30

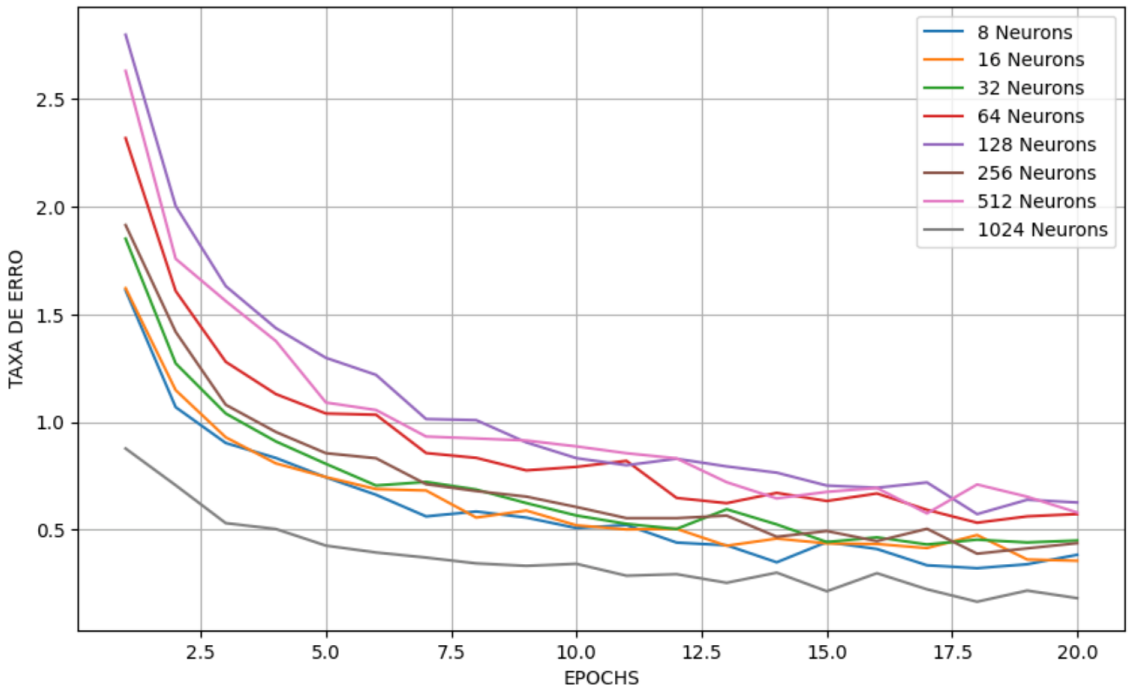


Figure 3. Taxa de erros com resolução 128x120

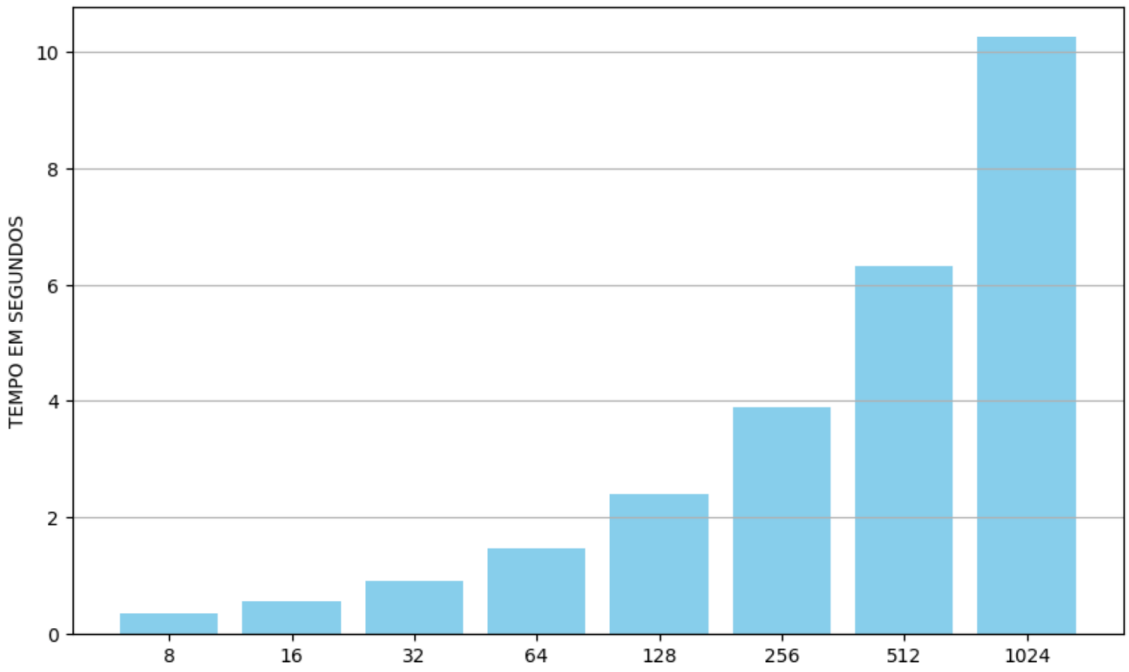


Figure 4. Tempo de treinamento com resolução 32x30

Foi feito uma média dos diversos resultados obtidos. Com as imagens em 32x30, foi feito uma média de 10 execuções diferentes. Já com as imagens em 128x120, foi feito 5 execuções diferentes. A seguir vem o tempo de execução de com cada resolução

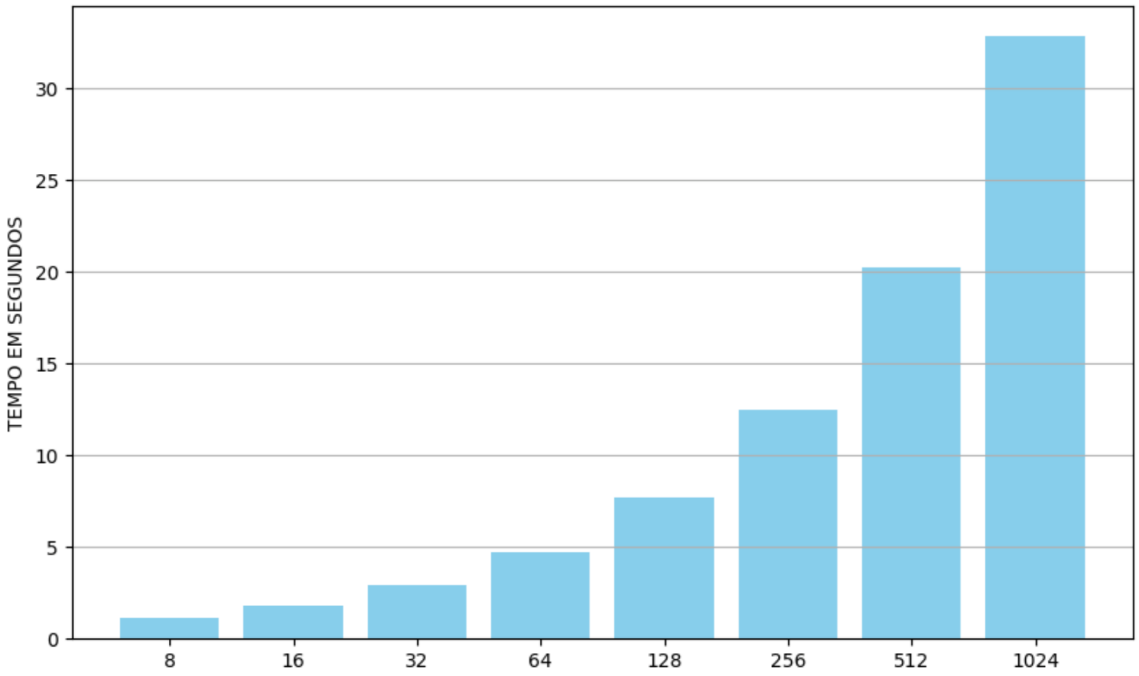


Figure 5. Tempo de treinamento com resolução 128x120

5. Conclusão

Portanto, esse trabalho conseguiu demonstrar que as técnicas de paralelização para redes neurais evoluíram significativamente desde 2013. Utilizando a GPU Tesla T4 no Google Colab, foi possível obter melhorias de desempenho substanciais em comparação com os resultados originais. As evoluções nas tecnologias de hardware e software definitivamente contribuíram para esses avanços.

6. Referencial Bibliográfico

References

[1] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[2] A. A. Huqqani, E. Schikuta, S. Ye, and P. Chen. Multicore and gpu parallelization of neural networks for face recognition. *Procedia Computer Science*, 18:349–358, 2013.

[3] D. B. Kirk and W. H. Wen-Mei. *Programming massively parallel processors: a hands-on approach*. Morgan kaufmann, 2016.